Bioinformatics

CS300

Prediction and Modeling Protein Structure

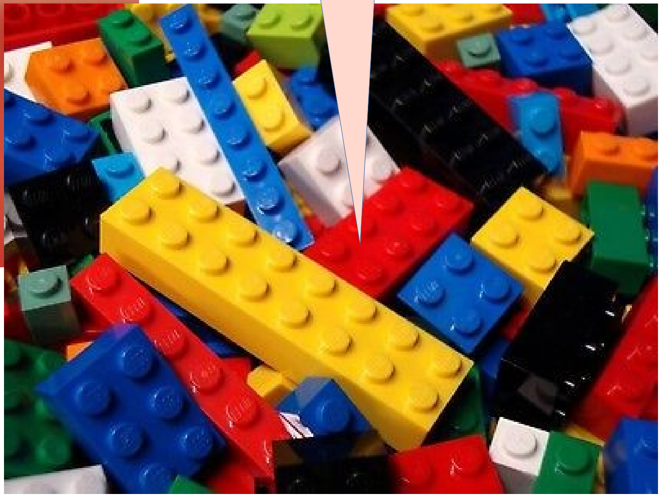
Week11, Deck 1

Fall 2022

Oliver BONHAM-CARTER

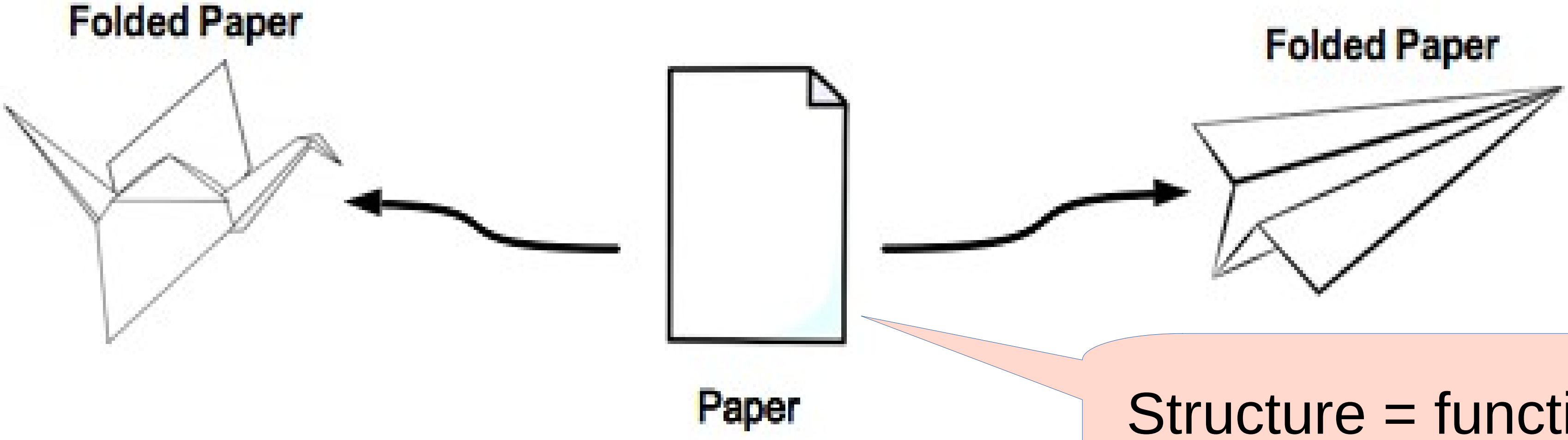
Properties From Combining Pieces

A cool living room made from Lego pieces!



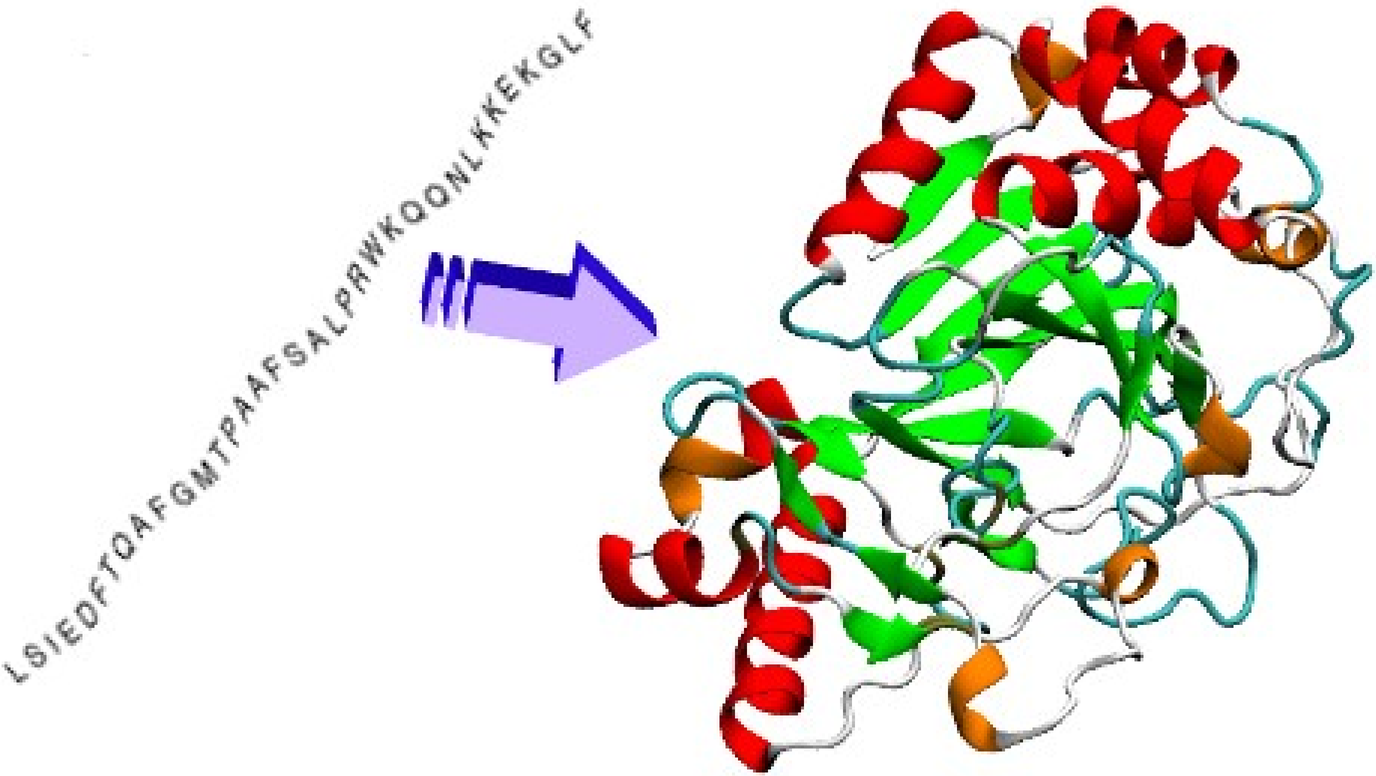
From these pieces?

Properties From Folding



Structure = function

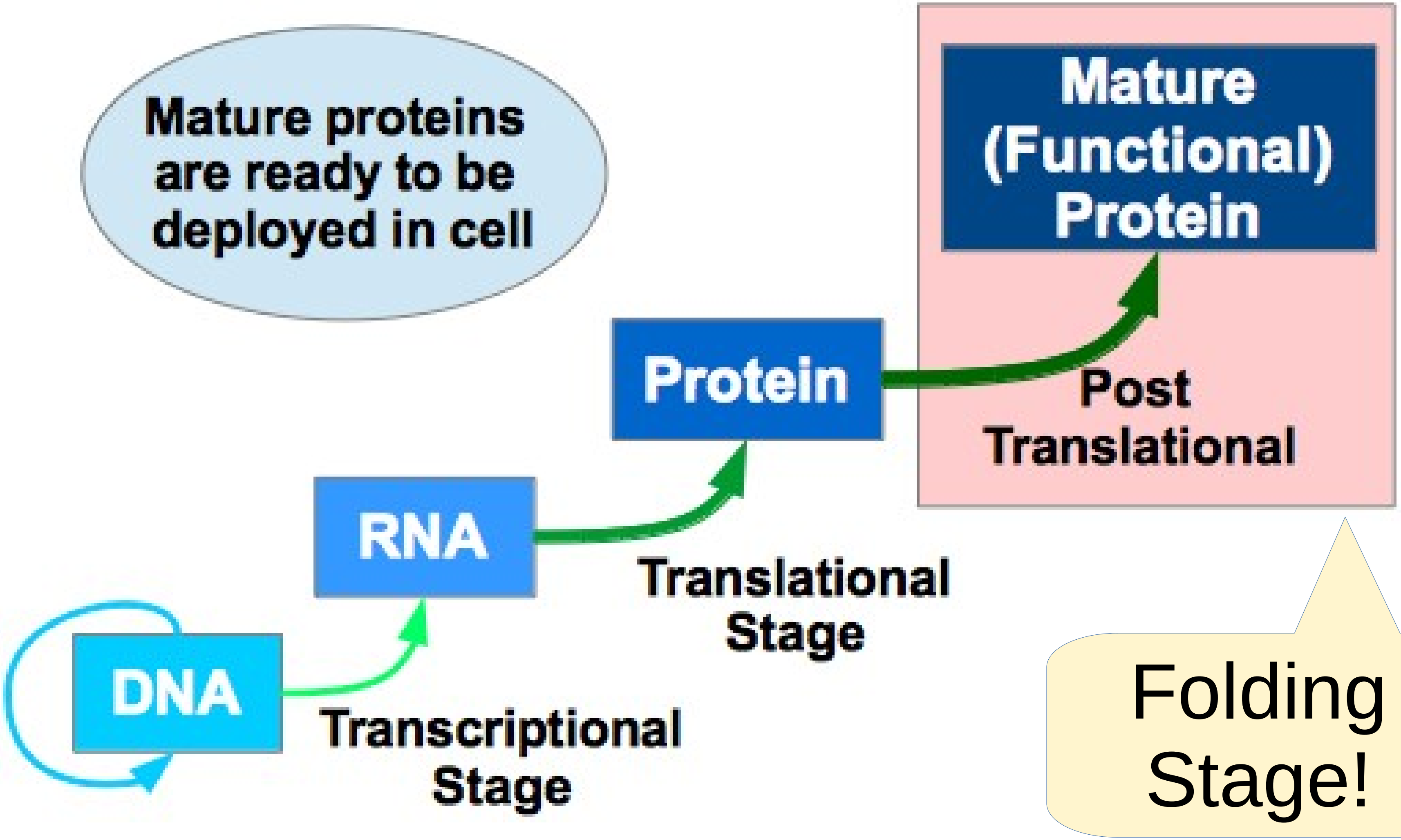
Protein Folding



- A protein sequence is a linear chain of amino acids produced by ribosomes during translation

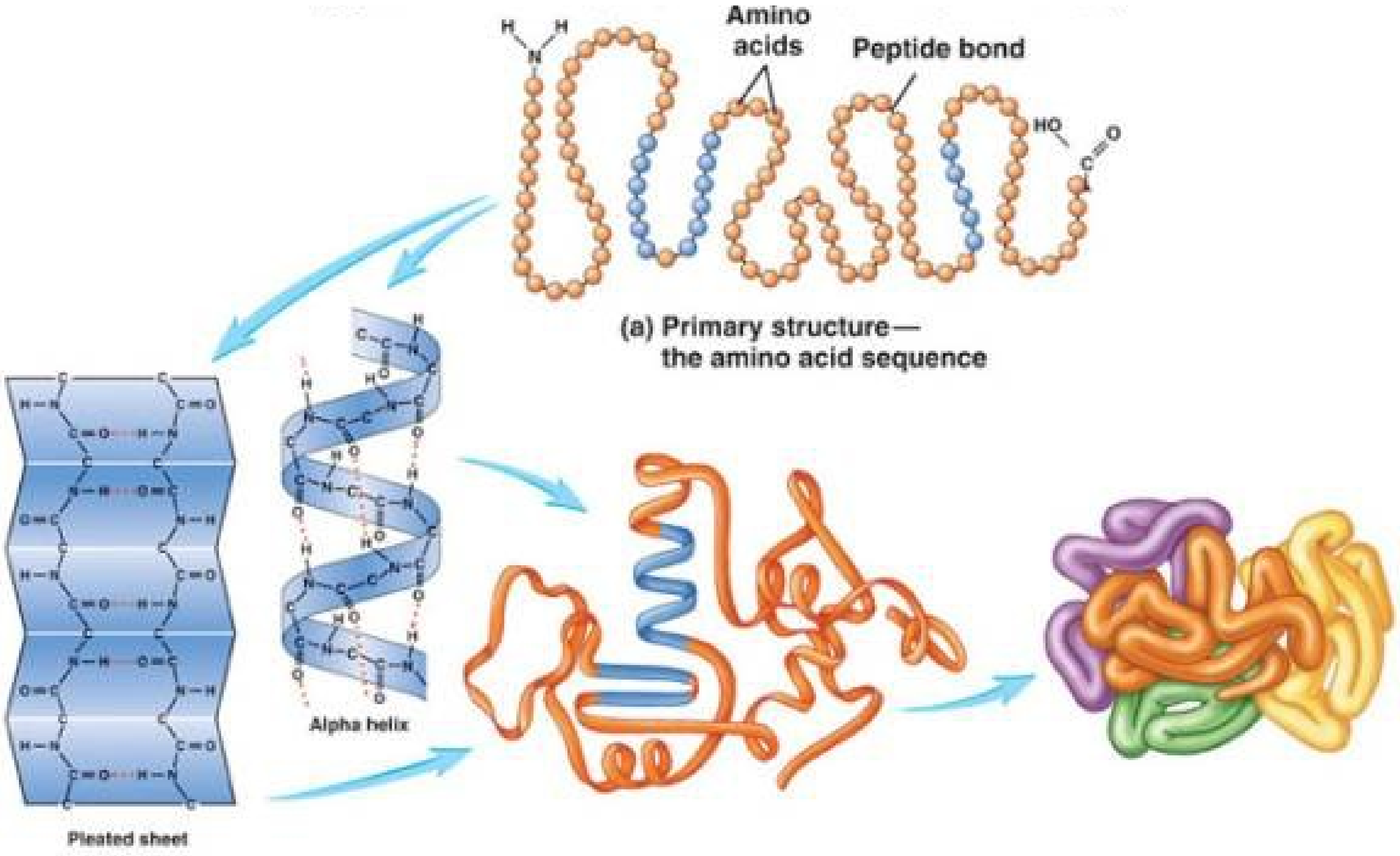
- A structure from folding, 3D state based on properties of amino acids and structure

Protein Folding and the Central Dogma of Biology


Mature proteins are ready to be deployed in cell
Mature (Functional) Protein
Protein
Post Translational
RNA
Translational Stage
DNA
Transcriptional Stage

Folding Stage!

Protein Folding: Four Stages



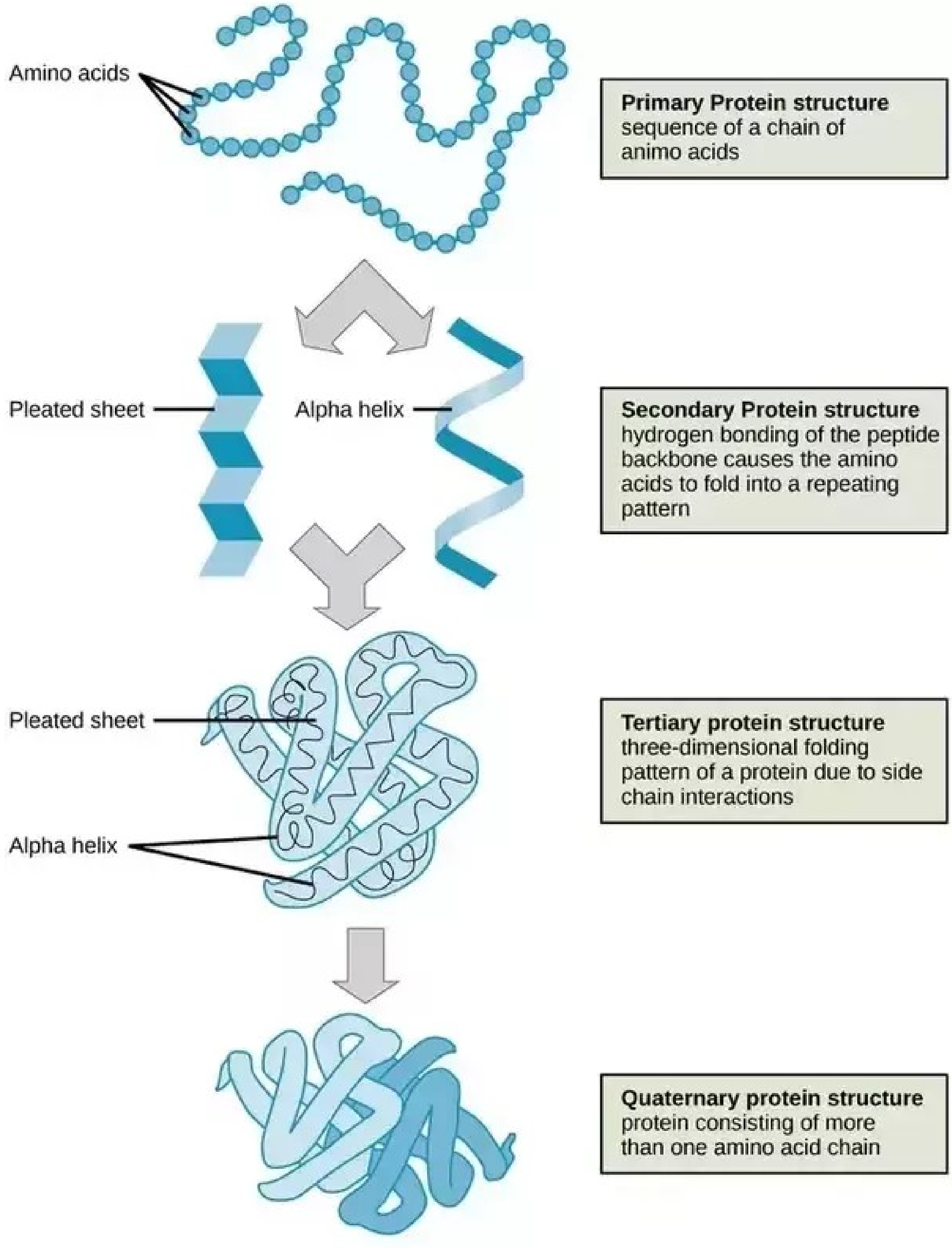
(a) Primary structure— the amino acid sequence

(b) Secondary structure with folding as a result of hydrogen bonding (*dotted red lines*)

(c) Tertiary structure with secondary folding caused by interactions within the polypeptide and its immediate environment

(d) Quaternary structure — the relationships between individual subunits

Protein Folding: Another View



Primary Protein structure
sequence of a chain of animo acids

Secondary Protein structure
hydrogen bonding of the peptide backbone causes the amino acids to fold into a repeating pattern

Tertiary protein structure
three-dimensional folding pattern of a protein due to side chain interactions

Quaternary protein structure
protein consisting of more than one amino acid chain

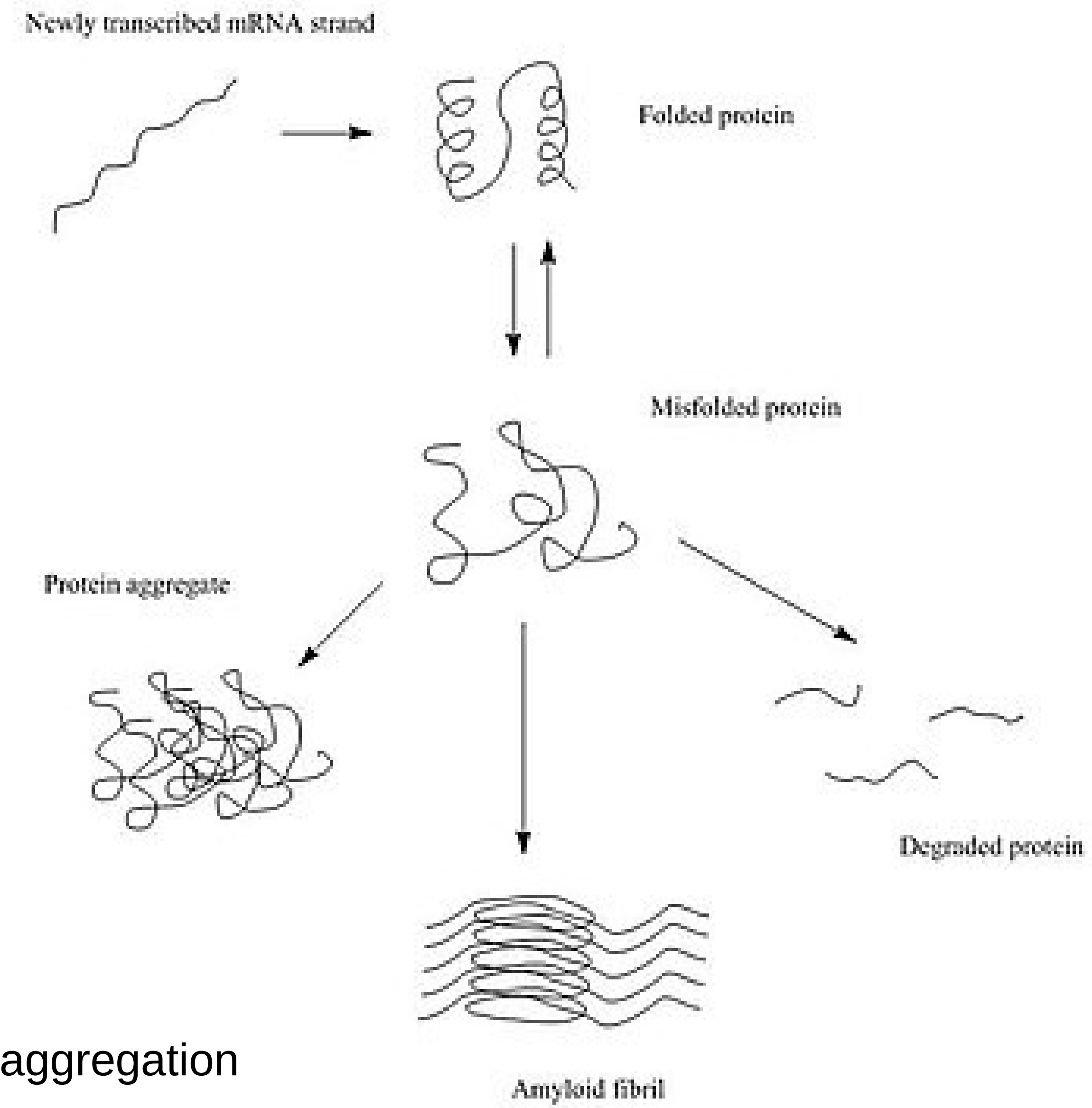
Supporting Videos

- Protein Folding (3 mins)
  - https://www.youtube.com/watch?v=yZ2aY5lxEGE
- What is a protein? (3D shape and function, 3 mins)
  - https://www.youtube.com/watch?v=qBRFIMcxZNM
- Protein folding simulation (3 mins)
  - https://www.youtube.com/watch?v=meNEUTn9Atg

Protein Folding - Applications

- **Protein must fold "*correctly*" to function "*correctly*"**

- Misfolded proteins
  - Accumulation (*clumping*) – Huntington's and Parkinson's disease
  - Tagged for degradation – emphysema, cystic fibrosis
    - Article: Pharmaceutical chaperones – therapies to fold mutated proteins to render them functional (placed in stabilized state)

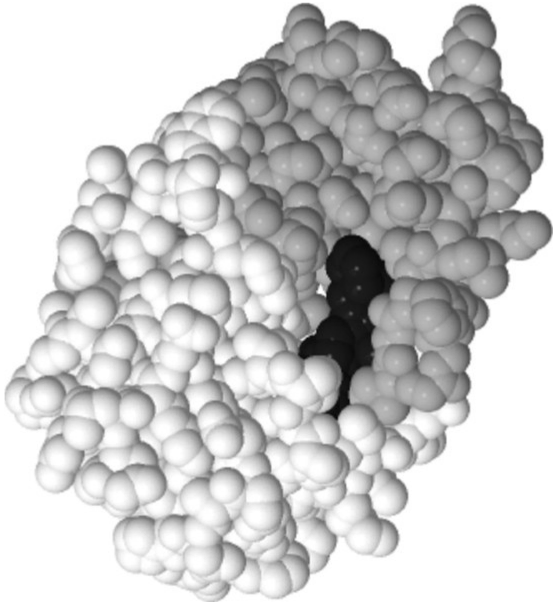


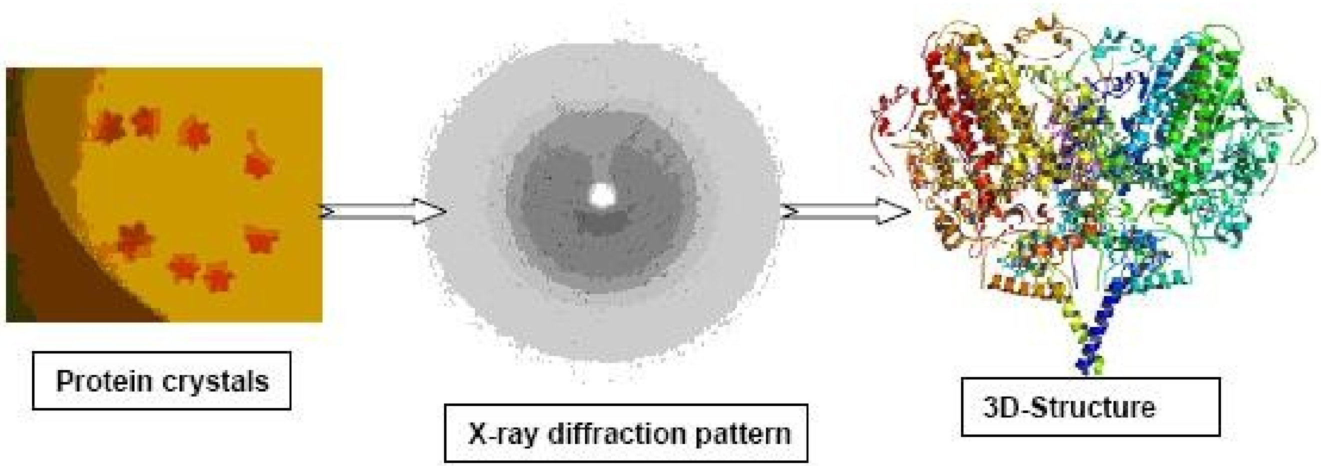
Ref: https://en.wikipedia.org/wiki/Protein_aggregation

Protein Folding - Applications

- Development of Antimicrobial Drugs: help to...
  - Be effective against the disease-causing agents
  - Be selectively toxic
  - kill or inhibit the microbe without harming the host

- Drugs Structures
  - Study 3-D structure (and function) of viral proteins
  - Design drugs to fit (dock to) to proteins and block functions
- Laboratory – challenging to predict 3-D structure





Genomics & Computational Structural Biology

Genomics (study)

- Determines the ordered sequence of nucleotides in a genome
- Determines/ assigns (predicted) functions to regions of nucleotides by annotations

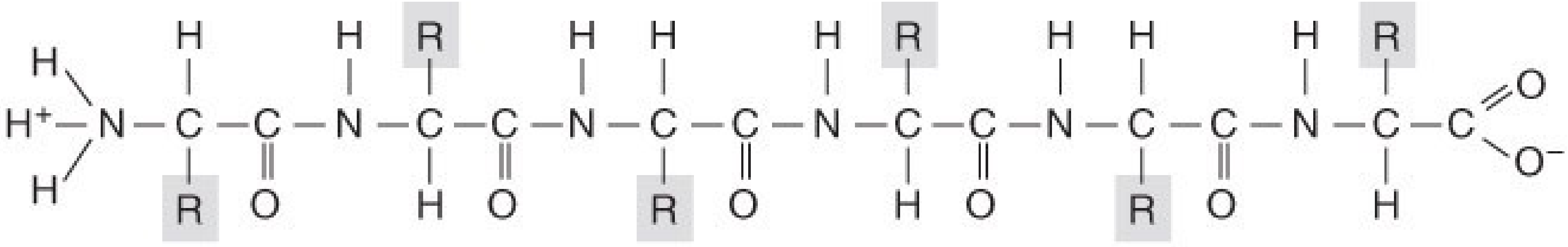
Computational Structural Biology (study)

- Takes predicted gene sequence for translation into primary amino acid sequence
- Predicts the 3-D protein structure based on the (primary) amino acid sequence
- Note: this step is very difficult because the number of possible outcomes to process and consider is enormous
- The study of structural rules and their contribution to the final mature protein.

Structural Rules for Protein Folding

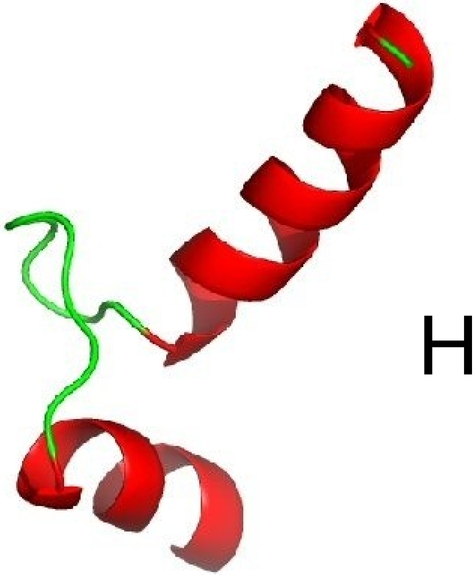
- Linus Pauling – Studied the limitations on protein folding
  - Nature of chemical bonds between amino acids
  - Bond angles
  - Rotation of atoms
  - Flexibility of side chains

Christian B. Anfisen – Studied the influence of thermodynamics of cellular environment



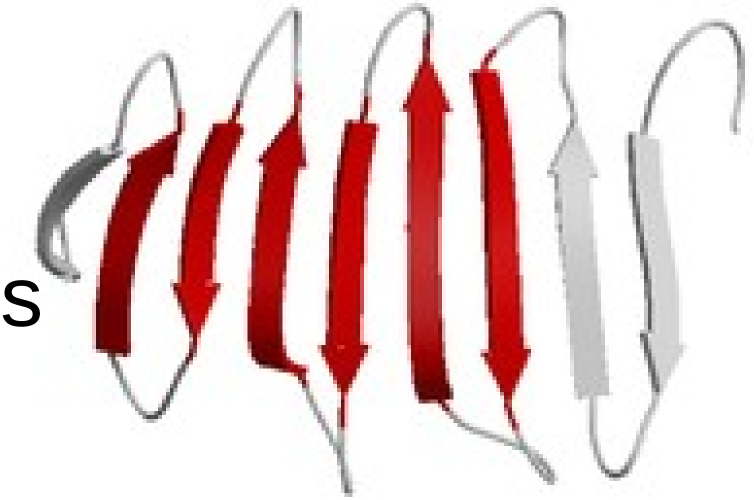
(A) Primary (1°) structure

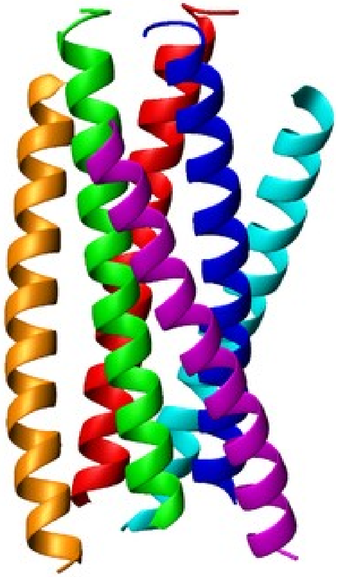
Parts of Protein (Structures)



Helices

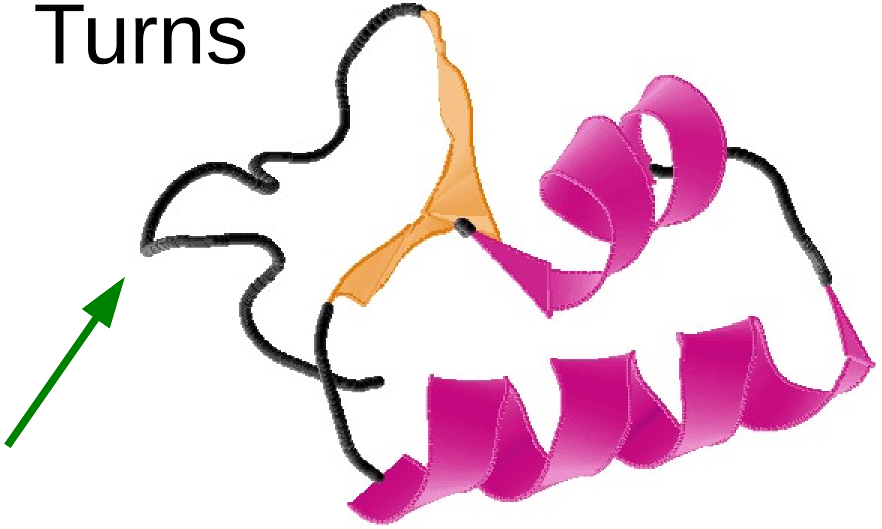
Sheets





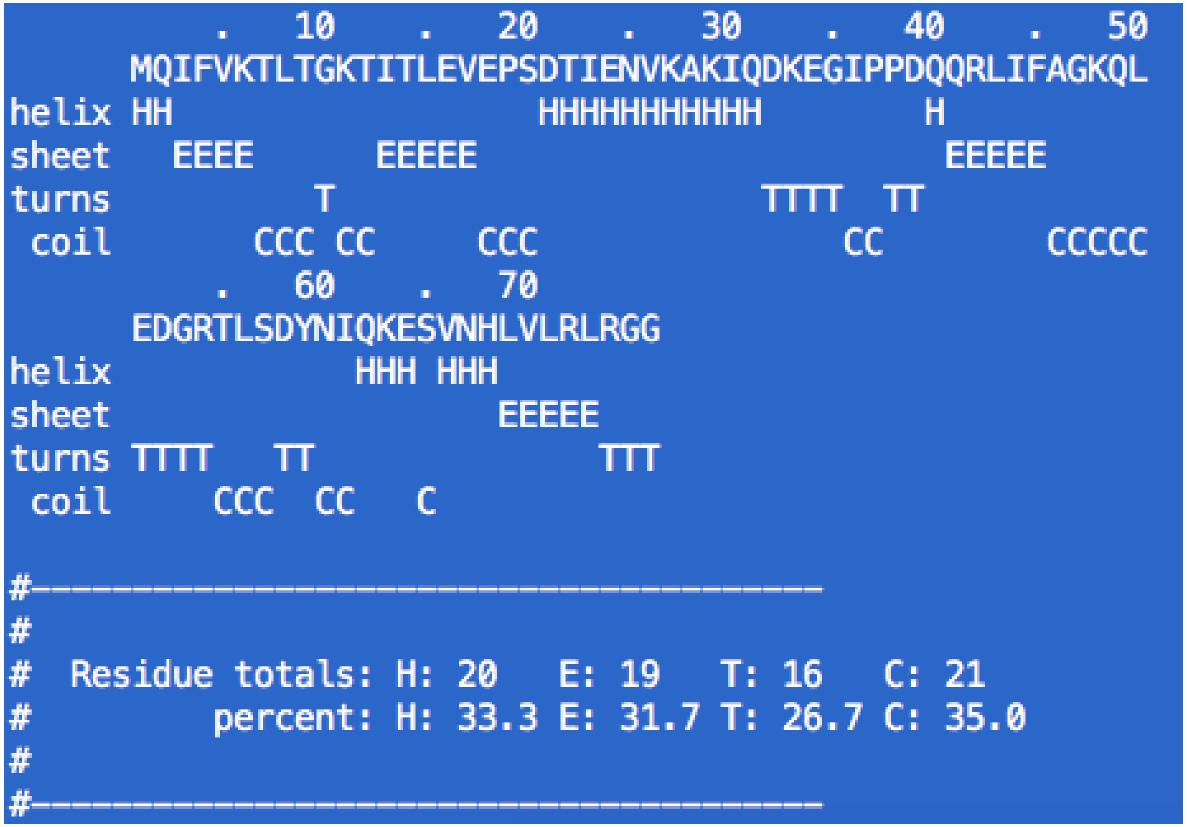
Coils

Turns



Protein Folding: An Idea of Structure

- **Garnier**: a text-based, command-line tool from EMBOSS
    - Input: protein sequence in fasta format
    - Output: a model of folding in text base
    - **Usage: garnier file.fasta**



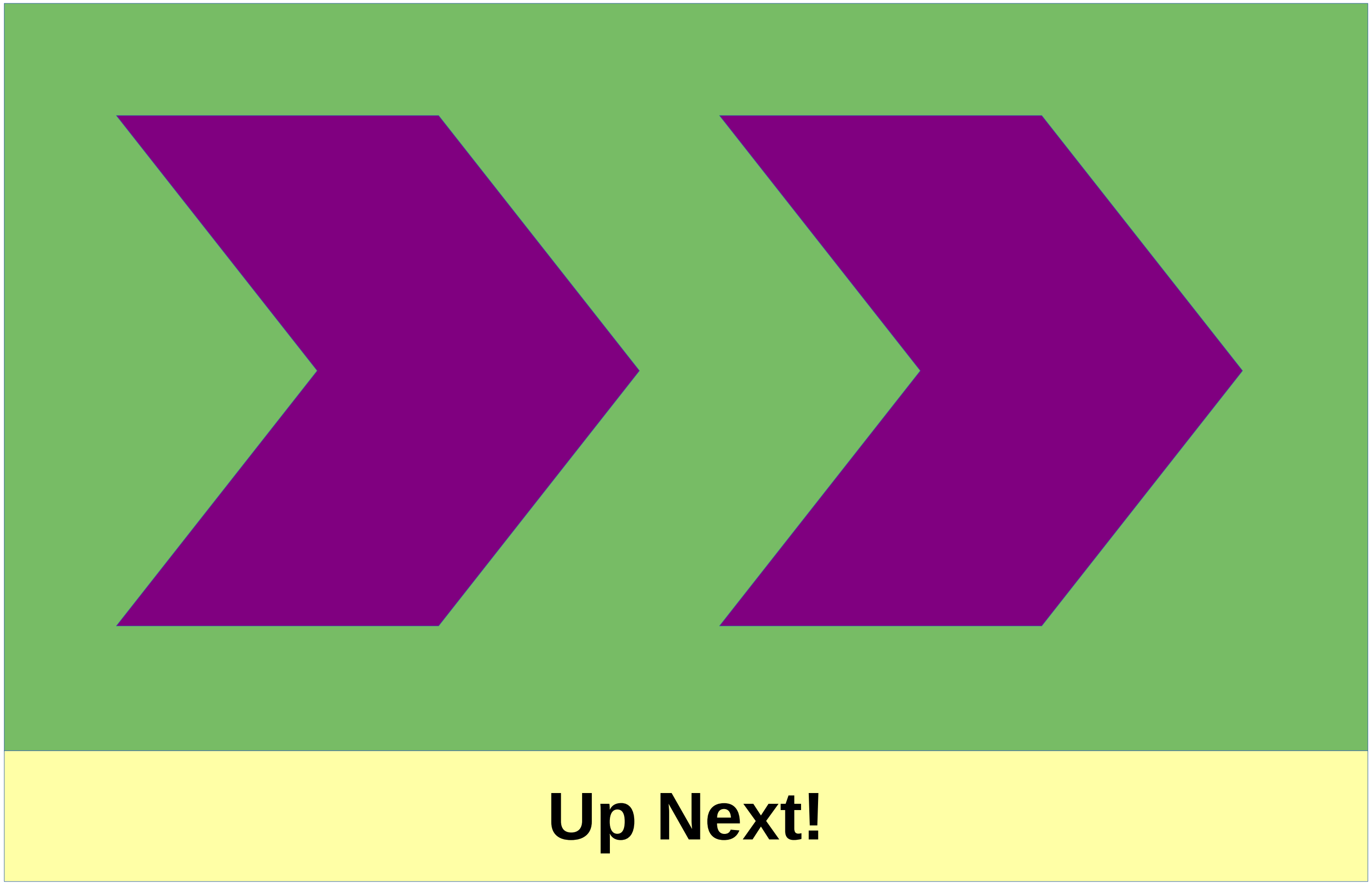
```
          .   10    .   20    .   30    .   40    .   50
      MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQL
helix HH                  HHHHHHHHHHH        H
sheet   EEEE      EEEEE                       EEEEE
turns          T                     TTTT  TT
 coil     CCC CC     CCC                  CC        CCCCC
          .   60    .   70
      EDGRTLSDYNIQKESVNHLVLRLRGG
helix            HHH HHH
sheet                   EEEEE
turns TTTT   TT              TTT
 coil     CCC  CC   C

#---------------------------------------
#
#  Residue totals: H: 20   E: 19   T: 16   C: 21
#         percent: H: 33.3 E: 31.7 T: 26.7 C: 35.0
#
#---------------------------------------
```

H: Helices, E: Sheets
T: Turns, C: Coils

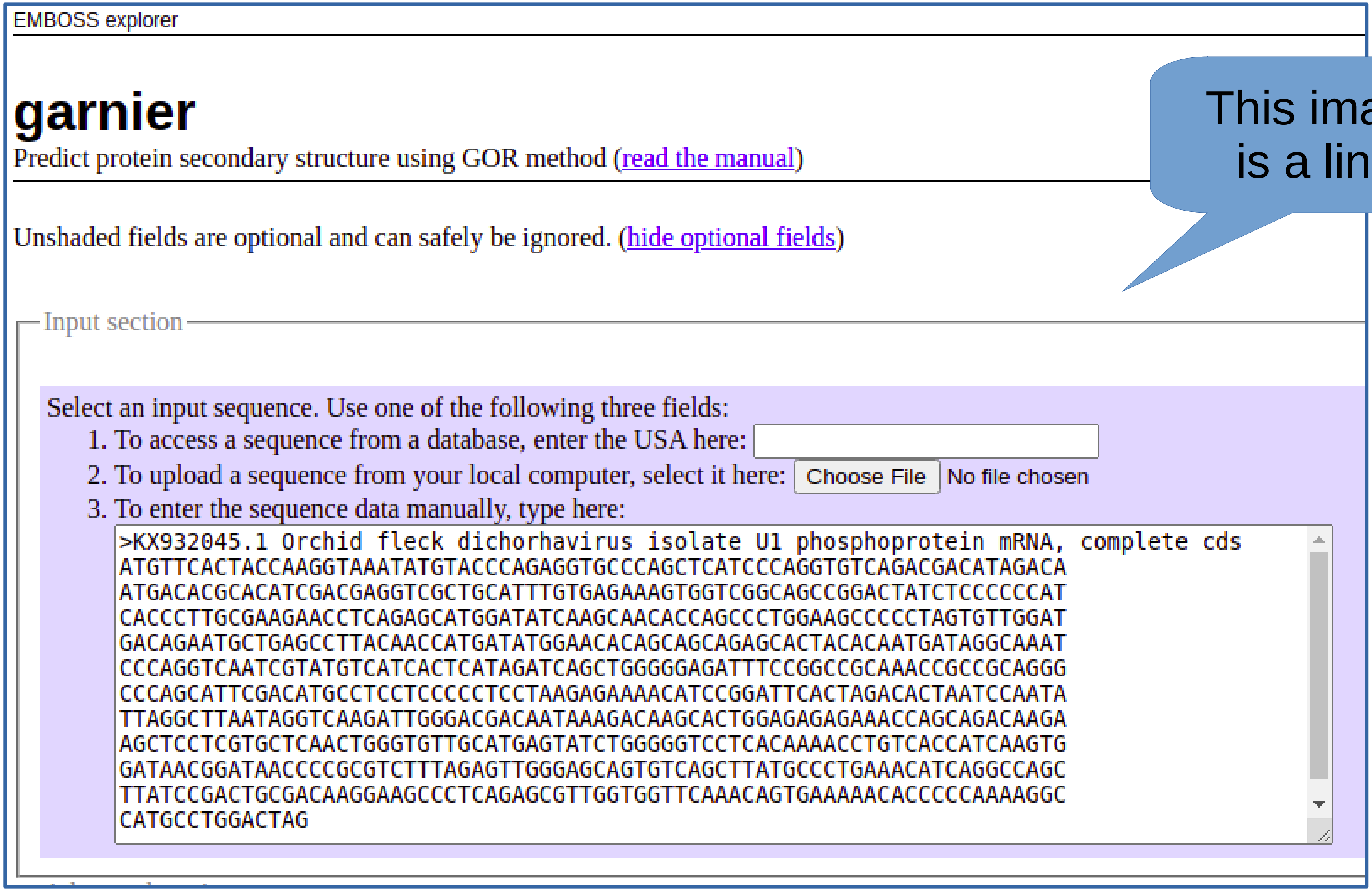
Ref: http://emboss.open-bio.org/rel/rel6/apps/garnier.html

Bring the Tool!



Up Next!

Protein Folding: Quick Solutions



EMBOSS explorer

garnier

Predict protein secondary structure using GOR method (read the manual)

Unshaded fields are optional and can safely be ignored. (hide optional fields)

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: Choose File No file chosen
3. To enter the sequence data manually, type here:

```
>KX932045.1 Orchid fleck dichorhavirus isolate U1 phosphoprotein mRNA, complete cds
ATGTTCACTACCAAGGTAAATATGTACCCAGAGGTGCCCAGCTCATCCCAGGTGTCAGACGACATAGACA
ATGACACGCACATCGACGAGGTCGCTGCATTTGTGAGAAAGTGGTCGGCAGCCGGACTATCTCCCCCCAT
CACCCTTGCGAAGAACCTCAGAGCATGGATATCAAGCAACACCAGCCCTGGAAGCCCCCTAGTGTTGGAT
GACAGAATGCTGAGCCTTACAACCATGATATGGAACACAGCAGCAGAGCACTACACAATGATAGGCAAAT
CCCAGGTCAATCGTATGTCATCACTCATAGATCAGCTGGGGGAGATTTCCGGCCGCAAACCGCCGCAGGG
CCCAGCATTCGACATGCCTCCTCCCCCTCCTAAGAGAAAACATCCGGATTCACTAGACACTAATCCAATA
TTAGGCTTAATAGGTCAAGATTGGGACGACAATAAAGACAAGCACTGGAGAGAGAAACCAGCAGACAAGA
AGCTCCTCGTGCTCAACTGGGTGTTGCATGAGTATCTGGGGGTCCTCACAAAACCTGTCACCATCAAGTG
GATAACGGATAACCCCGCGTCTTTAGAGTTGGGAGCAGTGTCAGCTTATGCCCTGAAACATCAGGCCAGC
TTATCCGACTGCGACAAGGAAGCCCTCAGAGCGTTGGTGGTTCAAACAGTGAAAAACACCCCCAAAAGGC
CATGCCTGGACTAG
```

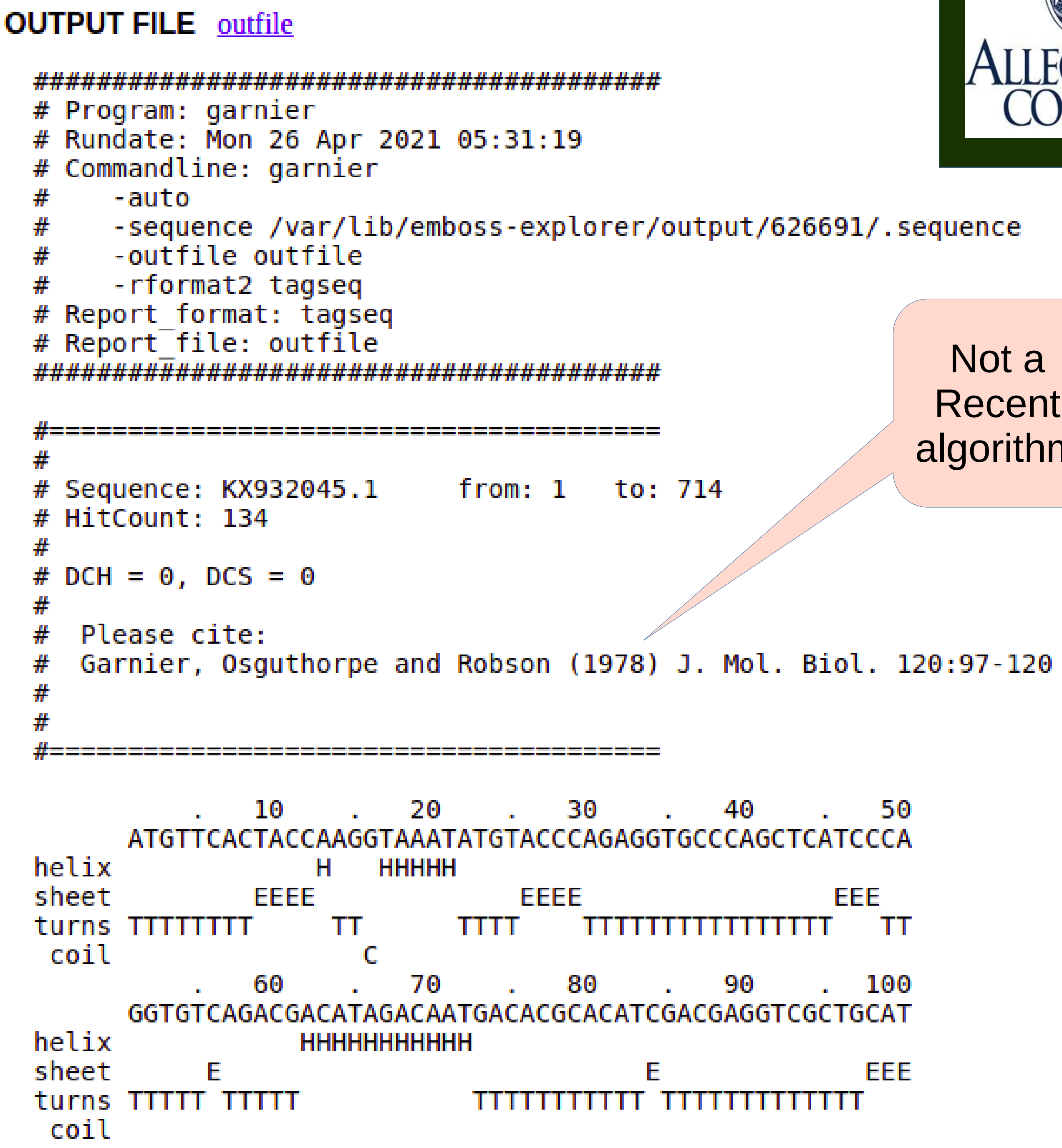
This image is a link!

https://www.bioinformatics.nl/cgi-bin/emboss/garnier

Article: https://www.sciencedirect.com/science/article/abs/pii/0022283678902978

Garnier Output (text)

H: Helices,
E: Sheets
T: Turns,
C: Coils



OUTPUT FILE outfile

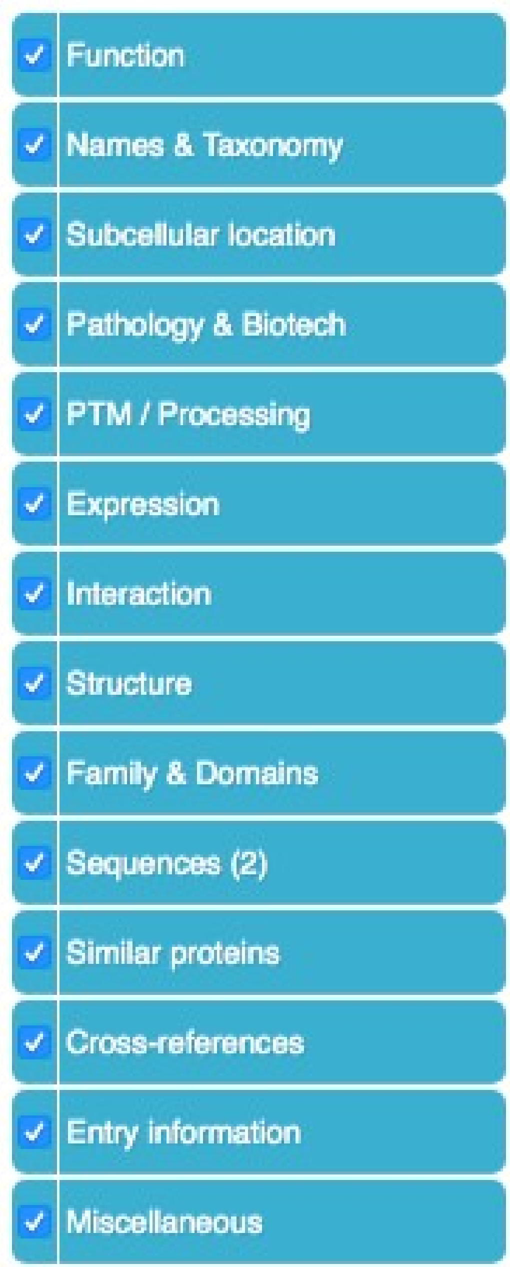
```
########################################
# Program: garnier
# Rundate: Mon 26 Apr 2021 05:31:19
# Commandline: garnier
#    -auto
#    -sequence /var/lib/emboss-explorer/output/626691/.sequence
#    -outfile outfile
#    -rformat2 tagseq
# Report_format: tagseq
# Report_file: outfile
########################################

#=======================================
#
# Sequence: KX932045.1     from: 1   to: 714
# HitCount: 134
#
# DCH = 0, DCS = 0
#
#  Please cite:
#  Garnier, Osguthorpe and Robson (1978) J. Mol. Biol. 120:97-120
#
#
#=======================================

          .   10    .   20    .   30    .   40    .   50
      ATGTTCACTACCAAGGTAAATATGTACCCAGAGGTGCCCAGCTCATCCCA
helix             H   HHHHH
sheet         EEEE             EEEE                EEE
turns TTTTTTTT     TT      TTTT    TTTTTTTTTTTTTTTT   TT
 coil                C
          .   60    .   70    .   80    .   90    .  100
      GGTGTCAGACGACATAGACAATGACACGCACATCGACGAGGTCGCTGCAT
helix            HHHHHHHHHHH
sheet      E                           E             EEE
turns TTTTT TTTTT           TTTTTTTTTTT TTTTTTTTTTTTT
 coil
```

Not a Recent algorithm

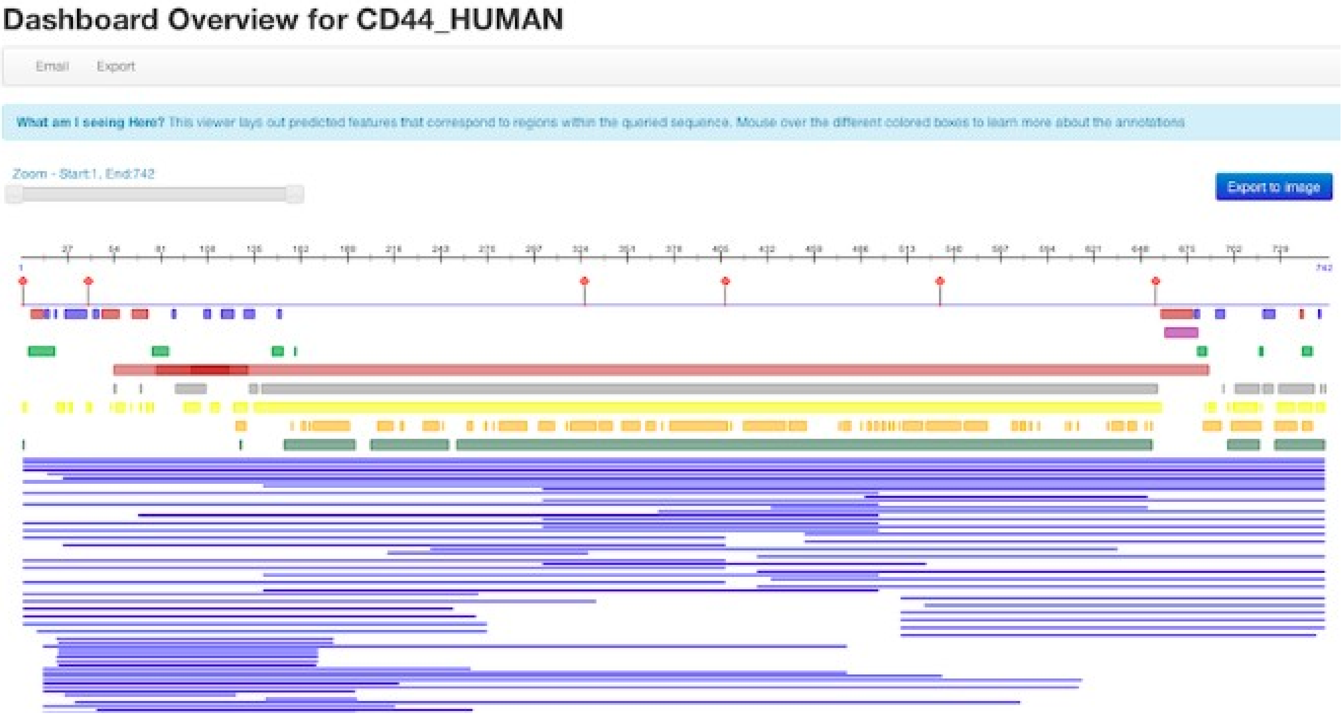
https://www.bioinformatics.nl/cgi-bin/emboss/garnier



Protein Information: The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

- Some protein study databases require information from UniProt to find protein samples.
- https://www.uniprot.org/

PREDICT PROTEIN



Dashboard Overview for CD44_HUMAN

What am I seeing Here? This viewer lays out predicted features that correspond to regions within the queried sequence. Mouse over the different colored boxes to learn more about the annotations

Zoom - Start:1, End:742

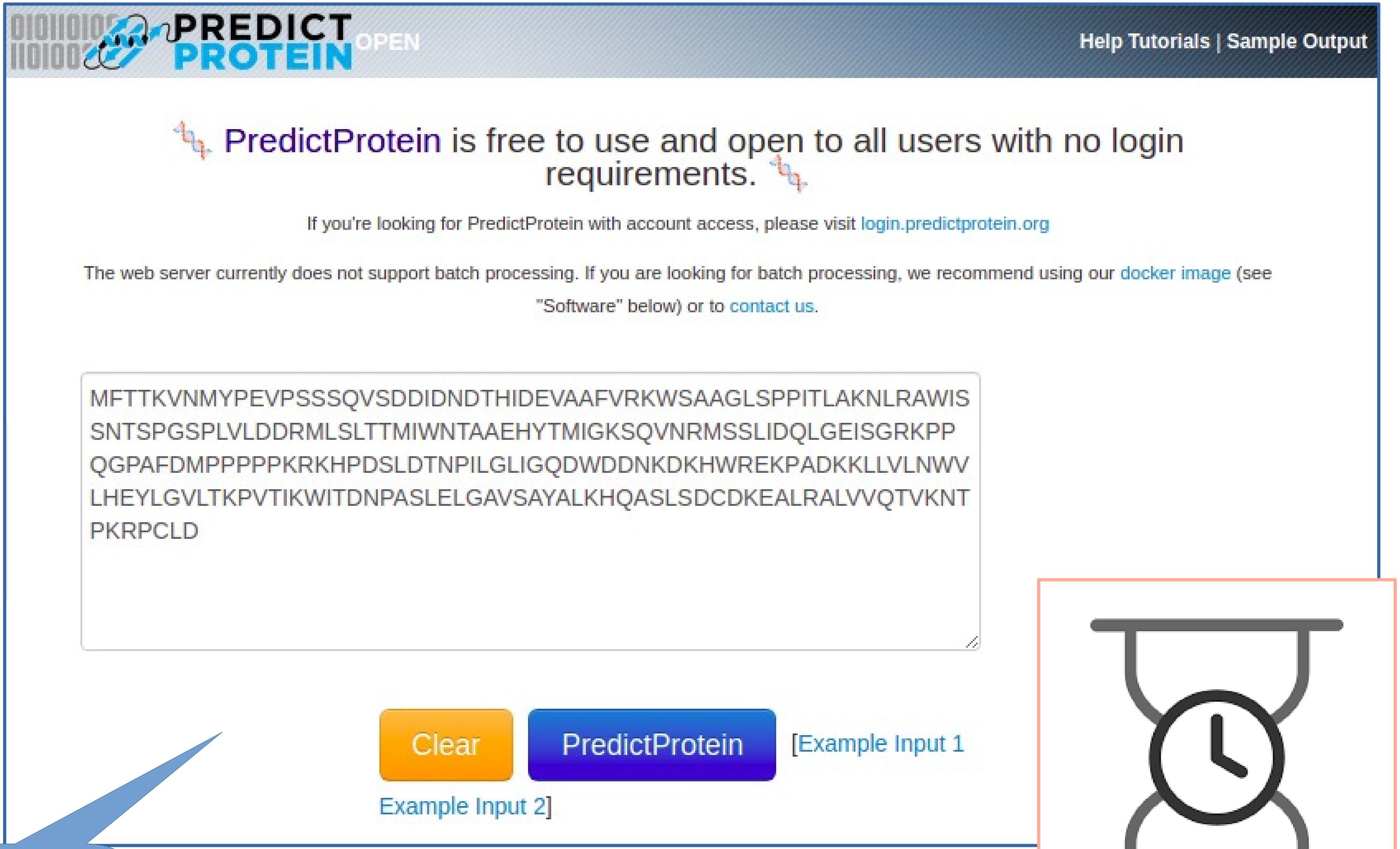
Structural Annotations of protein: prediction of protein function, e.g. assisting in the annotation of subcellular localization (LocTree, LocTree2, NLSpred), identifying protein-protein interaction sites (PPSites) and protein-DNA binding sites, and more.

- https://www.predictprotein.org/
- https://open.predictprotein.org/
- https://github.com/Rostlab/predictprotein-docker

Bring the Tool!

Up Next!

Protein Folding: Slower Solutions



PredictProtein is free to use and open to all users with no login requirements.

If you're looking for PredictProtein with account access, please visit login.predictprotein.org

The web server currently does not support batch processing. If you are looking for batch processing, we recommend using our docker image (see "Software" below) or to contact us.

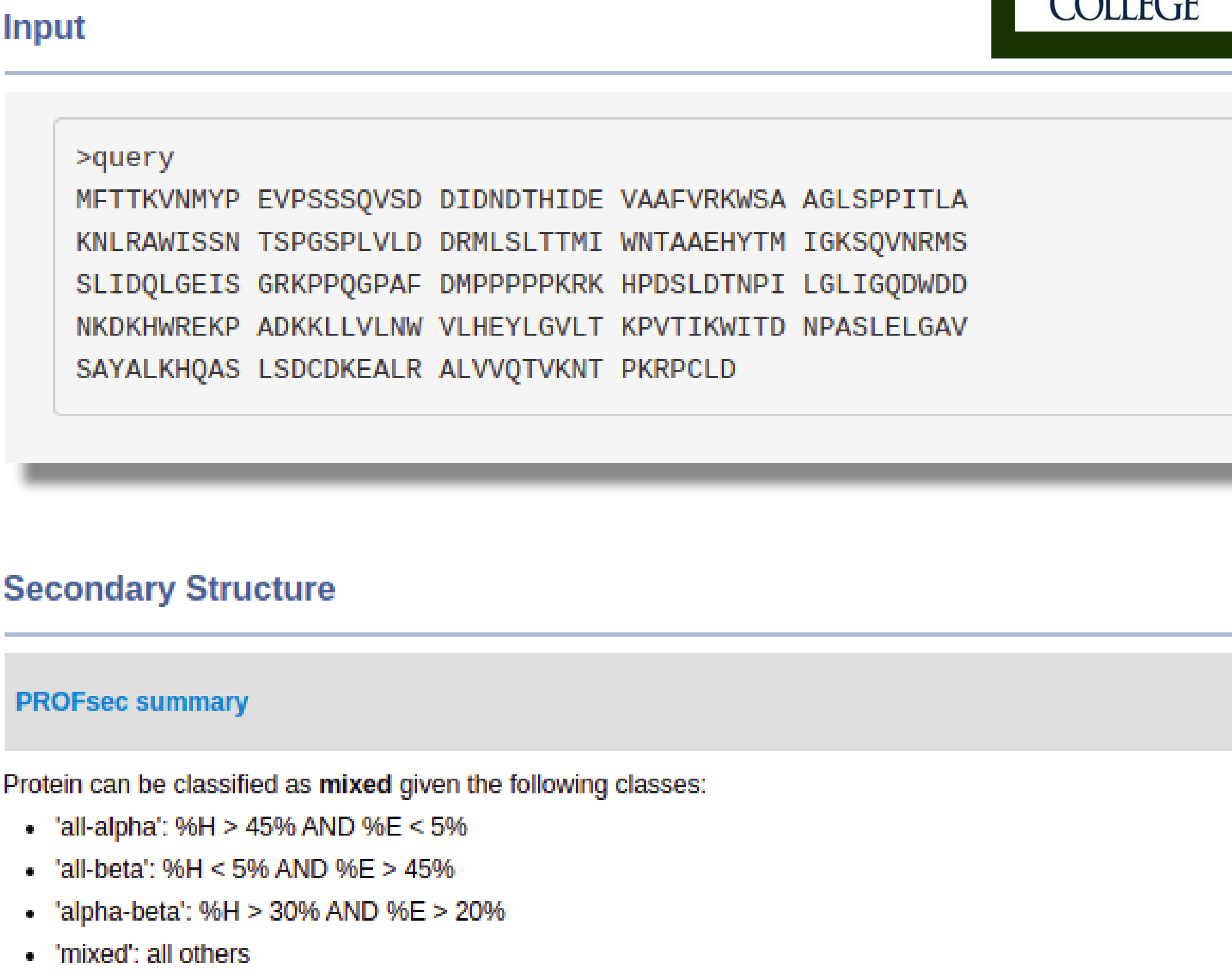
```
MFTTKVNMYPEVPSSSQVSDDIDNDTHIDEVAAFVRKWSAAGLSPPITLAKNLRAWIS
SNTSPGSPLVLDDRMLSLTTMIWNTAAEHYTMIGKSQVNRMSSLIDQLGEISGRKPP
QGPAFDMPPPPPKRKHPDSLDTNPILGLIGQDWDDNKDKHWREKPADKKLLVLNWV
LHEYLGVLTKPVTIKWITDNPASLELGAVSAYALKHQASLSDCDKEALRALVVQTVKNT
PKRPCLD
```

This image is a link!

- https://www.predictprotein.org/

Can take some time ...

Predict Protein output



Input

```
>query
MFTTKVNMYP EVPSSSQVSD DIDNDTHIDE VAAFVRKWSA AGLSPPITLA
KNLRAWISSN TSPGSPLVLD DRMLSLTTMI WNTAAEHYTM IGKSQVNRMS
SLIDQLGEIS GRKPPQGPAF DMPPPPPKRK HPDSLDTNPI LGLIGQDWDD
NKDKHWREKP ADKKLLVLNW VLHEYLGVLT KPVTIKWITD NPASLELGAV
SAYALKHQAS LSDCDKEALR ALVVQTVKNT PKRPCLD
```

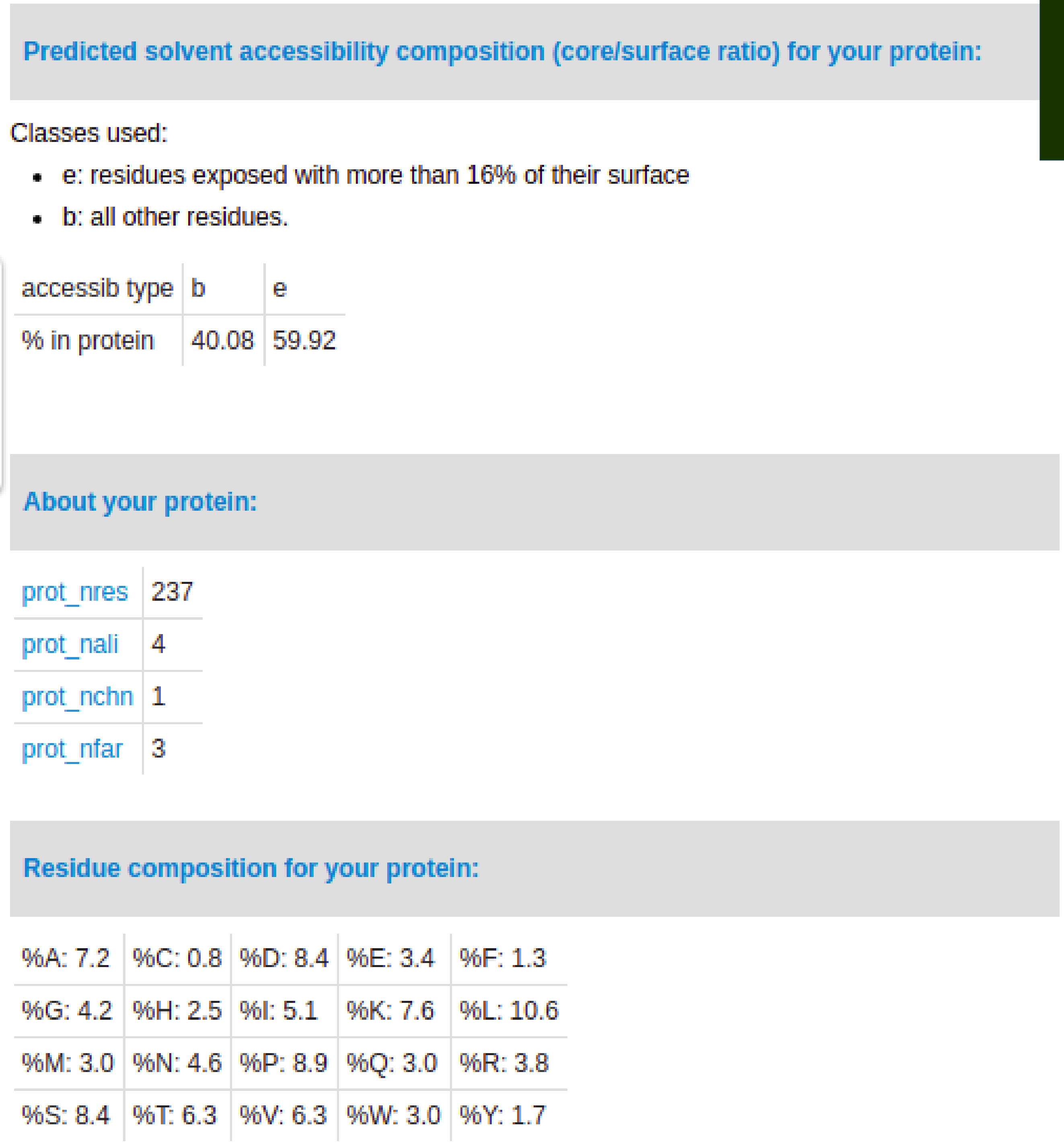
Secondary Structure

PROFsec summary

Protein can be classified as **mixed** given the following classes:

- 'all-alpha': %H > 45% AND %E < 5%
- 'all-beta': %H < 5% AND %E > 45%
- 'alpha-beta': %H > 30% AND %E > 20%
- 'mixed': all others

Predict Protein output



Predicted solvent accessibility composition (core/surface ratio) for your protein:

Classes used:

- e: residues exposed with more than 16% of their surface
- b: all other residues.

accessib type	b	e
% in protein	40.08	59.92

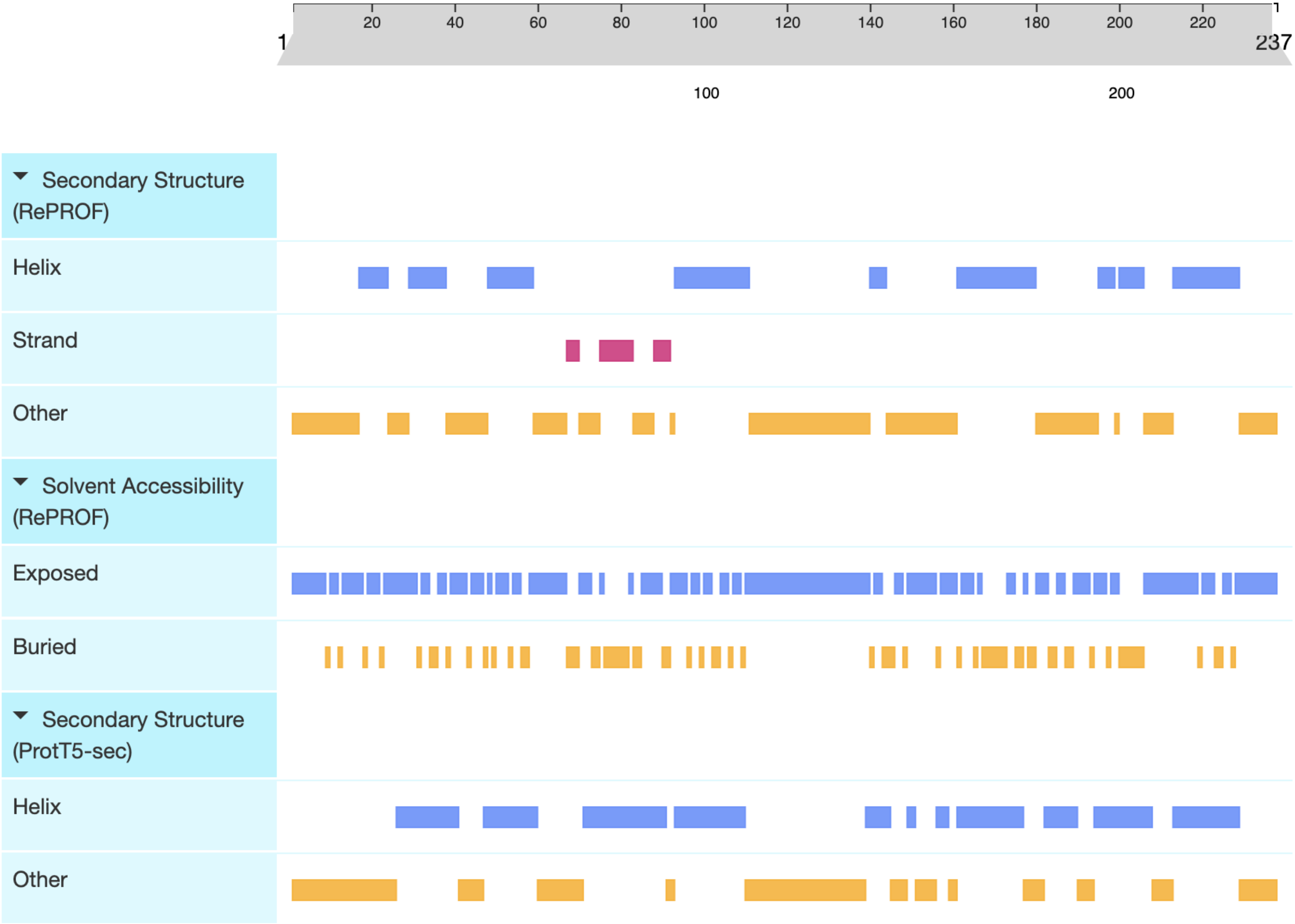
About your protein:

prot_nres	237
prot_nali	4
prot_nchn	1
prot_nfar	3

Residue composition for your protein:

%A: 7.2	%C: 0.8	%D: 8.4	%E: 3.4	%F: 1.3
%G: 4.2	%H: 2.5	%I: 5.1	%K: 7.6	%L: 10.6
%M: 3.0	%N: 4.6	%P: 8.9	%Q: 3.0	%R: 3.8
%S: 8.4	%T: 6.3	%V: 6.3	%W: 3.0	%Y: 1.7

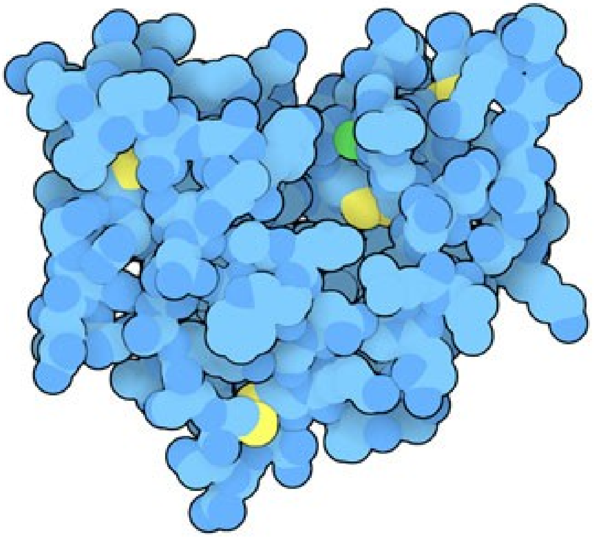
Predict Protein Output


1
20
40
60
80
100
120
140
160
180
200
220
237
100
200
Secondary Structure (RePROF)
Helix
Strand
Other
Solvent Accessibility (RePROF)
Exposed
Buried
Secondary Structure (ProtT5-sec)
Helix
Other

RCSB PDB PROTEIN DATA BANK

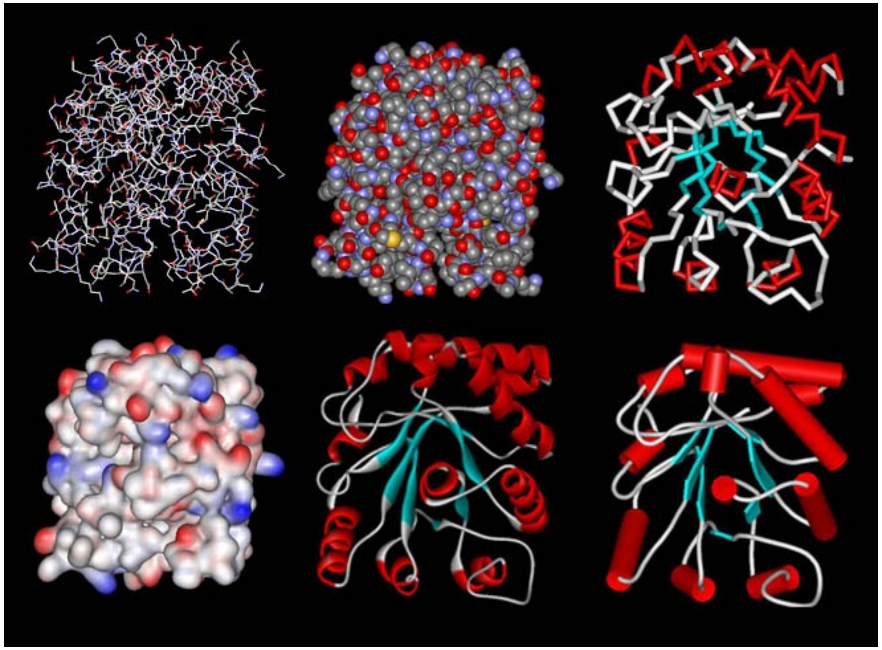
Protein archives: This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

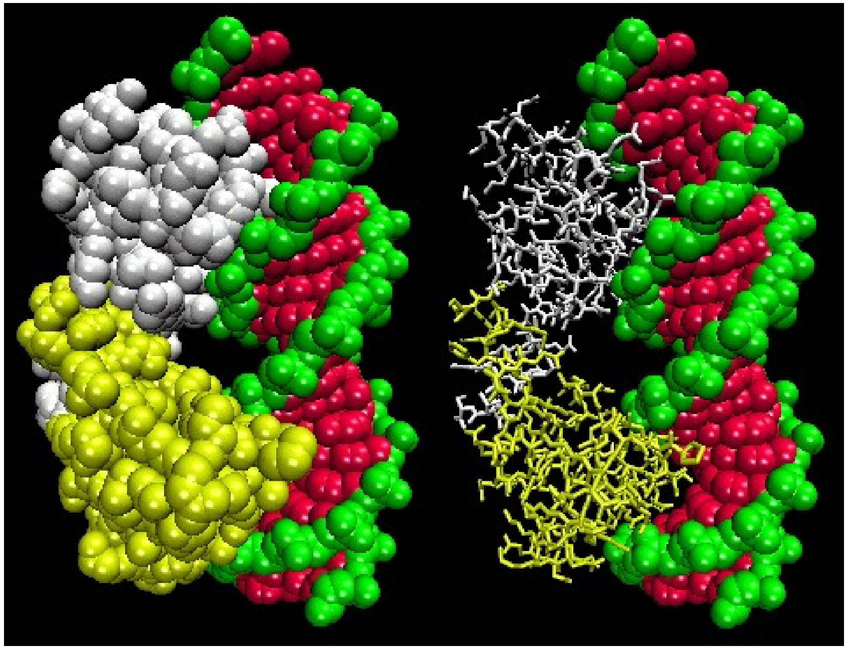
- http://www.rcsb.org/



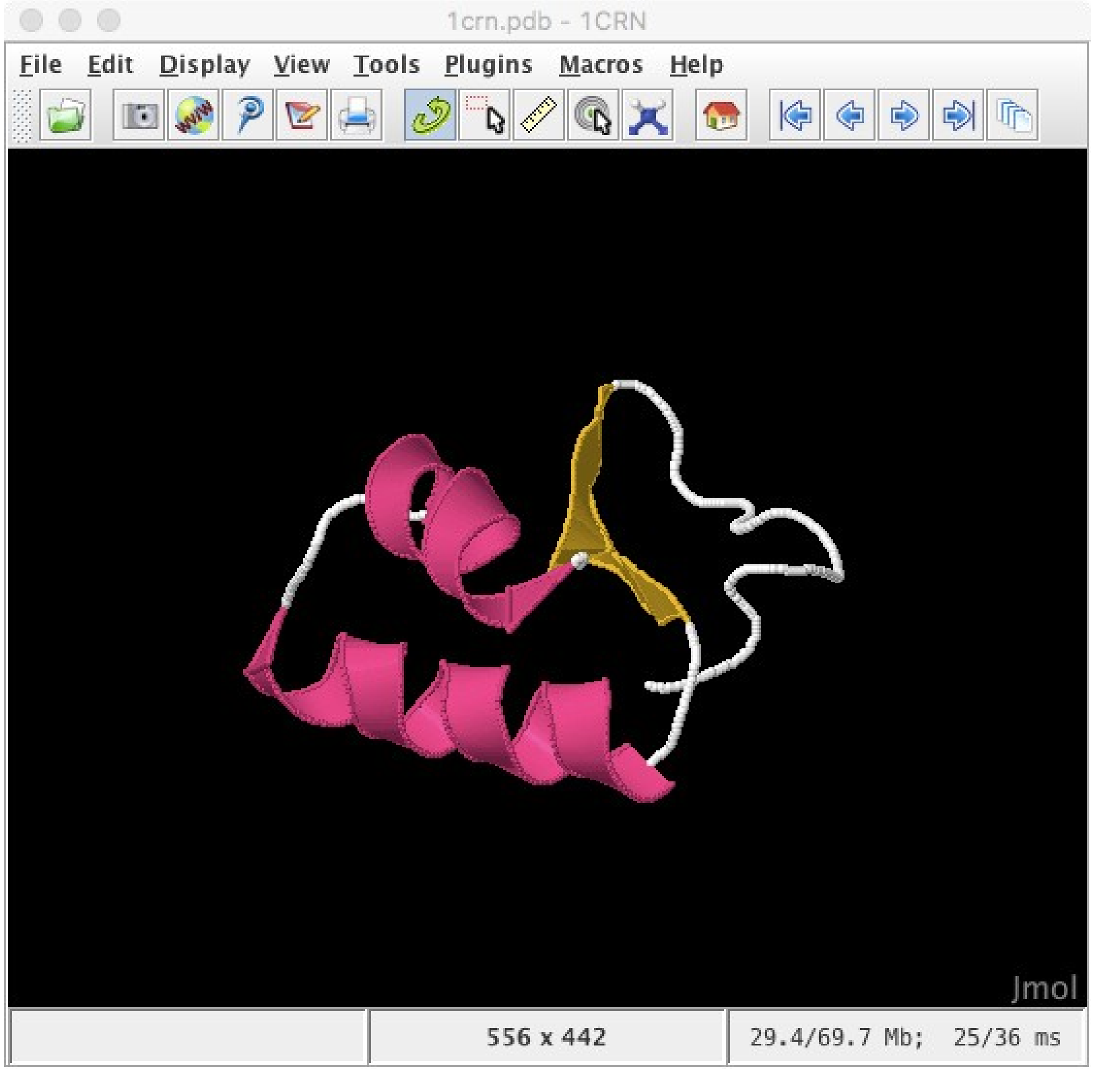
Protein DataBase (PDB)

- Database for 3-D structural data of large biological molecules
- https://www.rcsb.org/
- Data is viewable using jmol (local use) and with online tools.





Jmol: A (local) Graphical Viewer For Protein Sequences



556 x 442

29.4/69.7 Mb; 25/36 ms

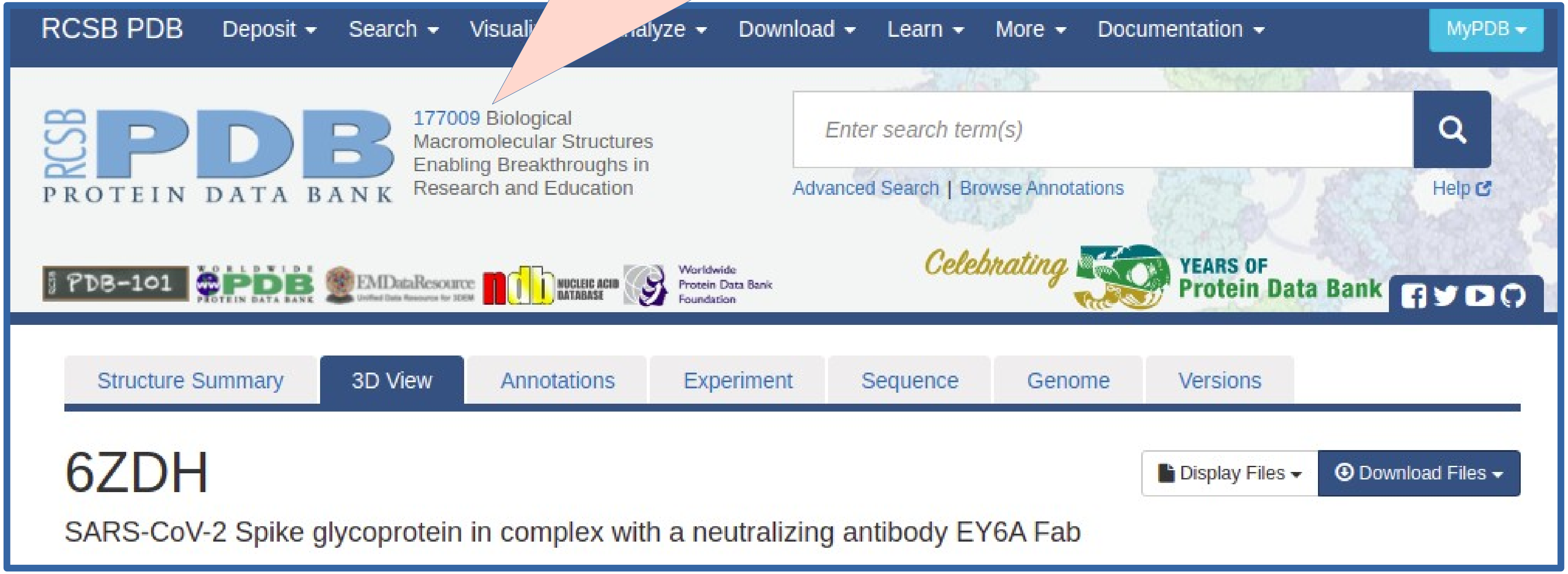
- Download:
  - http://jmol.sourceforge.net/
- Wiki:
  - http://wiki.jmol.org/index.php/Jmol_Application#Installing_Jmol_Application

Bring the Tool!

Up Next!

Protein Folding: Pre-Compiled Solutions

It takes a long time to virtually fold proteins. This data is already “folded” and you can view it as a folded protein structure.



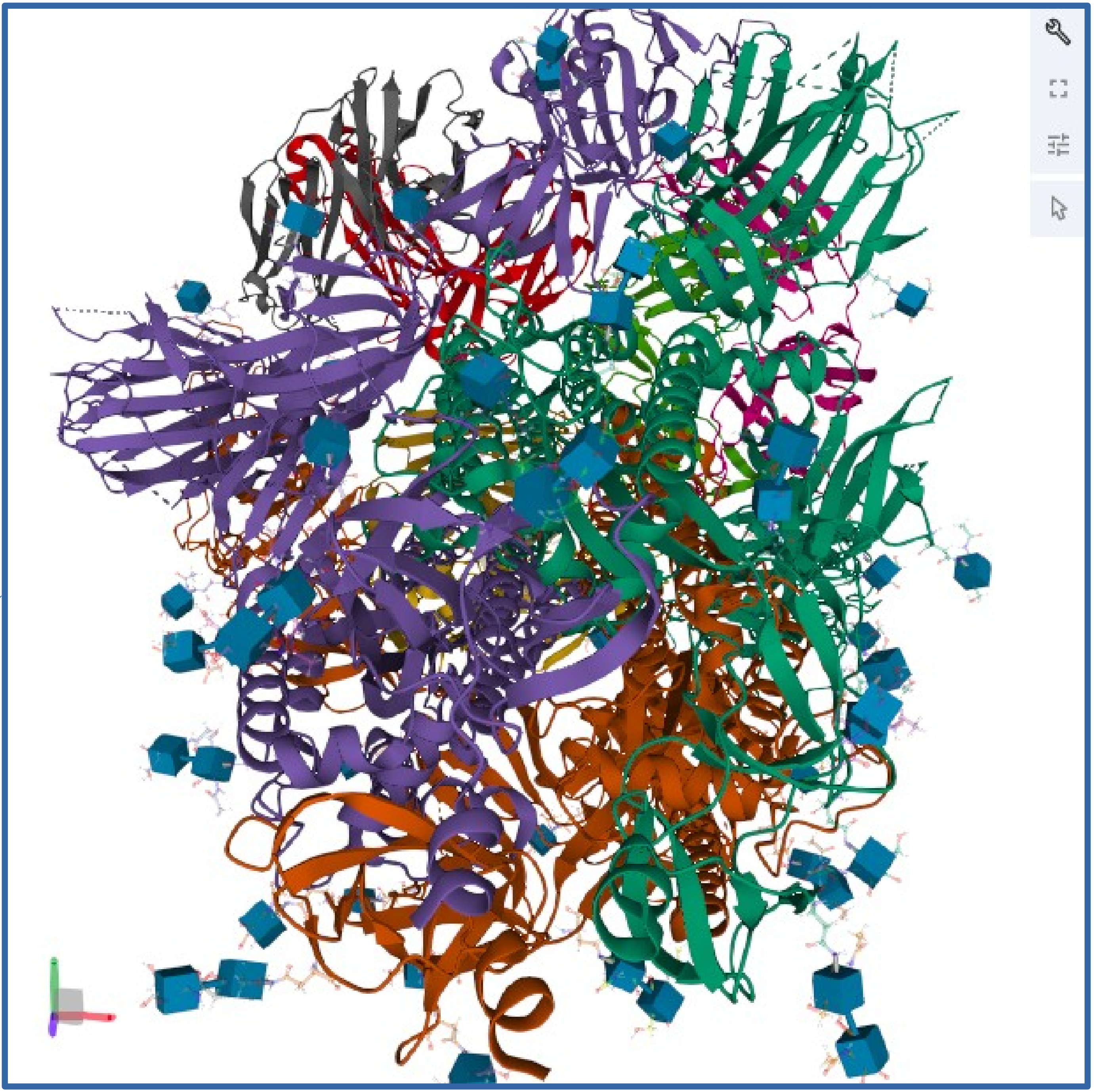
177009 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

6ZDH

SARS-CoV-2 Spike glycoprotein in complex with a neutralizing antibody EY6A Fab

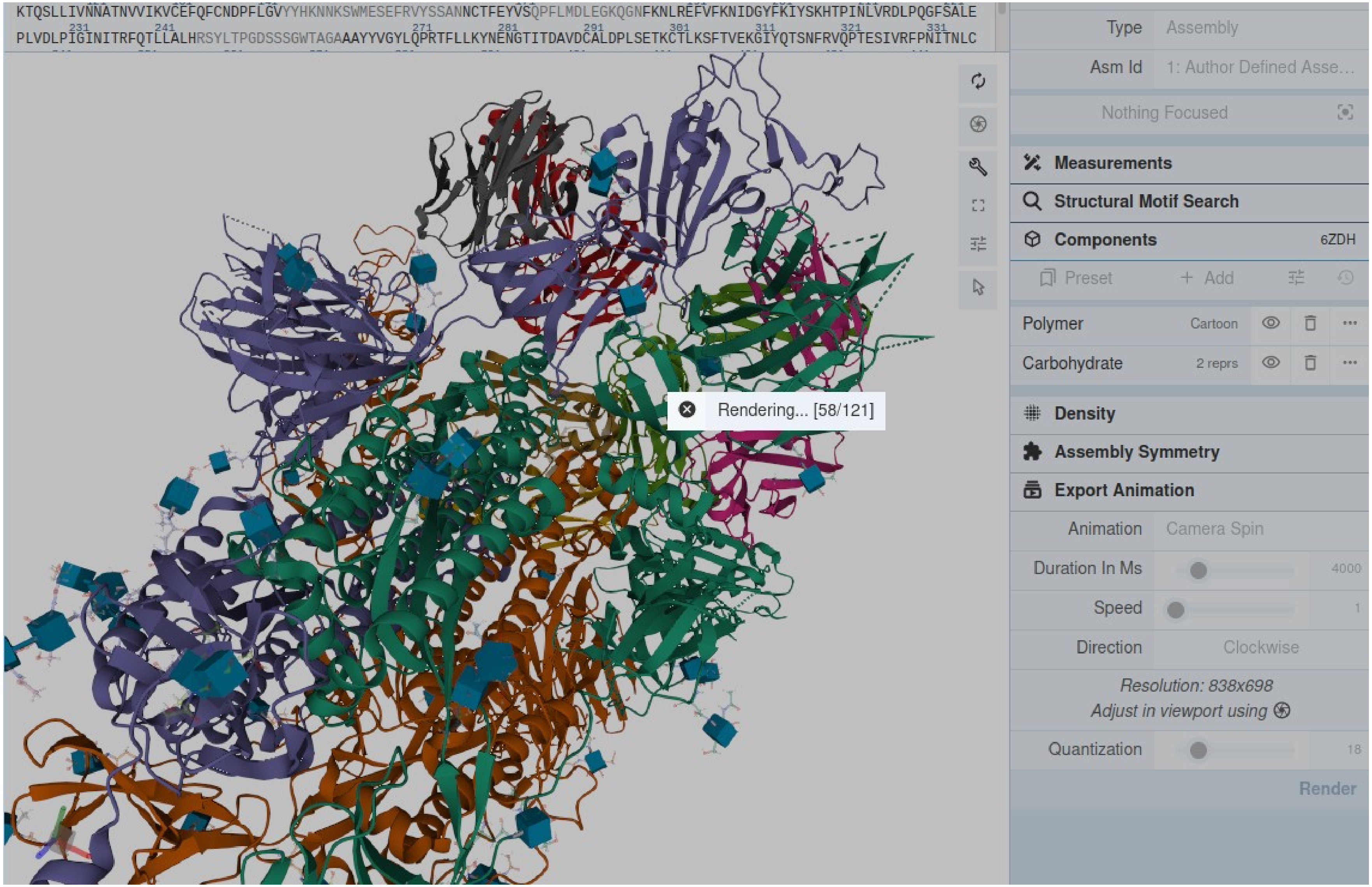
- http://www.rcsb.org/

RCSB Output



This image is a link!

Viewing Options To Animate



```
KTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALE
      231         241                            271       281       291        301        311        321        331
PLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLC
```

Rendering... [58/121]

Type: Assembly

Asm Id: 1: Author Defined Asse…

Nothing Focused

Measurements

Structural Motif Search

Components 6ZDH

Preset + Add

Polymer Cartoon

Carbohydrate 2 reprs

Density

Assembly Symmetry

Export Animation

Animation: Camera Spin

Duration In Ms: 4000

Speed: 1

Direction: Clockwise

Resolution: 838x698
Adjust in viewport using

Quantization: 18

Render

Save animation to a file